Kinematic data clustering for healthy knee gait characterization

Fatma Zgolli*†, Khadidja Henni*¶, Rim Haddad*, Amar Mitiche§, Youssef Ouakrim*¶,
Nicola Hagemeister¶, Pascal-André Vendittoli‖, Alexandre Fuentes**and Neila Mezghani*‖
*Laboratoire de recherche en Imagerie et orthopdie (LIO), Centre de Recherche LICEF, TELUQ university, Montreal, Canada
†École nationale d'Electronique et des Télécommunications de Sfax (ENET'Com), Sfax, Tunisie
‡Laboratoire Innov'Com, École Supérieur de communications de Tunis (Sup'Com), Tunis, Tunisie
§INRS-Centre Énergie, matériaux et télécommunications, Montreal, Quebec, Canada
¶Laboratoire LIO, École de technologie supérieure, Centre de Recherche du CHUM, Montreal, Quebec, Canada
‖Centre de recherche de l'Hopital de Maisonneuve-Rosemont, Montreal, Quebec, Canada
**Centre du genou EMOVI, Quebec, Canada

***Abstract*—The purpose of this study is to investigate data clustering to determine representative patterns in three-dimensional (3D) knee kinematic data measurements. Kinematic data are high-dimensional vectors to describe the temporal variations of the three fundamental angles of knee rotation during a walking cycle, namely the abduction/adduction angle, with respect to the frontal plane, the flexion/extension angle, with respect to the sagittal plane, and internal/external angle, with respect to the transverse plane. To offset the curse of dimensionality, inherent to high dimensional data pattern analysis, the method reduces dimensionality by isometric mapping without affecting information content. The data thus simplified is then clustered by the DBSCAN algorithm. The method has been tested on a large database of 165 healthy knee kinematic data measurements. Clusters are validated in terms of the silhouette index, the Dunn index, and connectivity. Results show that a two-cluster characterization of the kinematic knee data in each plane is quite effective. A further clinical investigation shows that the men and women knee patterns are balanced between the two clusters and, for 80% of participants, the right and left knees are in the same cluster.**

I. Introduction

The interpretation of knee kinematic during locomotion is a subject of increasing interest in biomechanics research. The purpose is to evaluate e knee function objectively [1] so as to understand pathological knee alterations [2]. A characterization of knee kinematic data by a few representative patterns can inform on an individual's locomotion function [3] and thus assist in the diagnosis of normal gait, also called asymptomatic gait. The kinematic data of the knee describe the three angles between the tibia and femur in 3D space corresponding to flexion/extension in the sagittal plane, abduction/adduction in the frontal plane and internal/external rotation in the transverse plane. These data suffer from significant variability and also from the curse of dimensionality [4] due to their high dimensionality (Fig. 1). Most studies have used simple descriptions of the pathological classes, such as the mean of available gait data, or locally determined information, and have followed with clustering, including hierarchical [5], c-means [6], and fuzzy clustering [7]. In general, summarizing data by local information and average values, has led to poor interpretations. Some studies [8] have sought better interpretations by using global information in the form of kinematic curves.

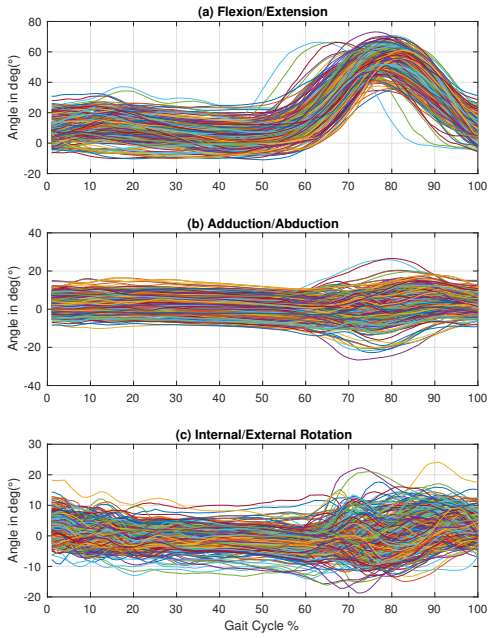

Fig. 1. 3D knee kinematic curves. Each curve represents a subject from the database : (a) Flexion/Extension, (b) Adduction/Abduction and (c) Internal/external rotation.

In this paper, we investigate density-based spatial clustering, namely the BDSCAN algorithm, of knee kinematic measurements curves to extract representative kinematic data curve that characterizes healthy gait of locomotion. Prior to clustering, and to offset the curse of dimensionality [4], the dimension of the data space is significantly reduced, while preserving the data descriptive content, by a nonlinear isometric mapping which preserves geodesic distances between clustered data. The method has been tested for each of the three measurement planes separately, namely the sagittal, frontal, and transverse

978-1-5386-6709-5/18/$31.00 ©2018 IEEE

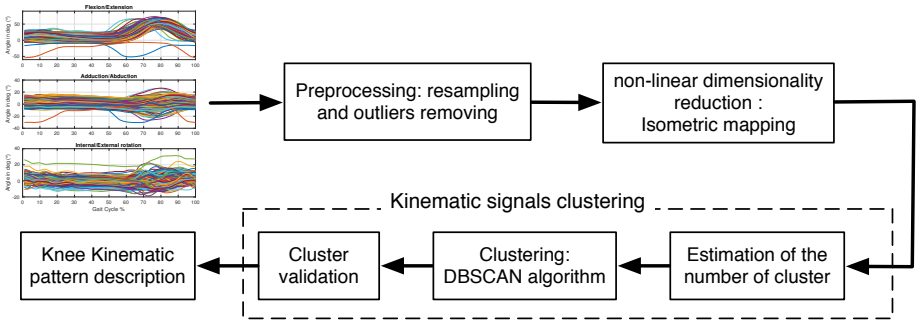

Fig. 2. Bloc diagram of the proposed knee kinematic clustering method.

planes. Cluster divisions of the data are evaluated using the silhouette index, the Dunn index, and connectivity. Results show that a two-cluster characterization of the kinematic knee data in each plane is quite effective. A further clinical investigation shows that the men and women knee patterns are balanced between the two clusters and for 80% of participants, the right and left knees are in the same cluster.

The remainder of this paper is organized as follows: Section II describes the method in its main functional steps, including dimensionality reduction, clustering, and cluster validation. The experimental results and a discussion are provided in Section III. Section IV contains a conclusion.

II. Method

The functional diagram of the knee kinematic data clustering method in this study is illustrated in Fig. 2. Following data collection and preprocessing (Section II-A), the proposed framework consists of three main steps: the first consists of nonlinear dimensionality reduction using an isometric mapping (Section II-B). This is followed by clustering of the kinematic data of reduced dimension, which includes the estimation of the number of clusters (Section II-C), clustering proper (Section II-D) and cluster validation (Section II-E). The resulting clusters are described based on clinical interpretation (Section III).

A. Data collection and preprocessing

3D knee kinematic data measurements consist of vectors that describe the temporal variation, during a full gait cycle of locomotion, of the three fundamental angles of knee rotation, i.e., the knee angles with respect to the sagittal, frontal, and transverse planes (Fig. 1). The data collection was performed using a state-of-the-art KneeKG acquisition system [9]. For each participant, the positional angles are recorded during about 45 sec on a treadmill. A total of 90 subjects (49 females and 41 males) were recruited: 83 from The Hospital Maisonneuve-Rosemont (HMR) and 7 from the Laboratoire de recherche en imagerie et orthopdie (LIO). Kinematics was analyzed on both knees of the 83 HMR subjects, and on one knee of the 7 LIO subjects. 8 of the HMR analyses (6 right knees and 2 left knees of different subjects) were excluded because of calibration errors or instability of the KneeKG giving measurements from a total of 165 knees. A mean kinematic pattern per subject was obtained by averaging the 15 most repeatable gait cycles. The knee rotation curves, defining the motion of the tibia relative to the femur, were then normalized from 1 to 100% of the average gait cycle. Data normalization was followed by outliers removal.

B. Dimensionality reduction

Dimensionality reduction has been performed using isometric mapping (IsoMap), a nonlinear dimensionality reduction method based on spectral theory. The main idea of IsoMap consists of performing a multidimensional scaling in the geodesic space in order to find the low-dimensional mapping that preserves the pairwise distances. The geodesic distance, which is the shortest path along the curved surface of the manifold, is approximately based on the nearest neighborhood graph [10].

C. Estimation of the number of clusters

The number of clusters is determined based on two criteria: the Bayesian information criterion and the intra-cluster variation using the Elbow method.

1) Bayesien information criterion (BIC): The Bayesian information criterion (BIC) is given by the general expression Kass and Wasserman [11]:

$$BIC = L(\theta) - \frac{1}{2} m \log n \tag{1}$$

where $L(\theta)$ is the log-likelihood function of data θ according to each model, m is the number of clusters and n is the size of the dataset? In our case θ corresponds, in each plane, to the knee kinematic data. The knee point in the BIC curve, which corresponds to the local maximum with highest probability is used to approximate the number k of clusters.

2) The Elbow method: This method searches the optimal number of clusters by minimizing the total intra-cluster variation (or the total within-cluster sum of the square) [12]:

$$\min(\sum_{k=1}^{K} W(C_k)) \tag{2}$$

where C_k is the k th cluster and $W(C_k)$ is the within-cluster variation.

D. Clustering using DBSCAN algorithm

DBSCAN (density-based spatial clustering of applications with noise) is the pioneer of the density-based clustering family [13], which considers clusters as dense regions separated by low-density regions. DBSCAN is able to detect clusters of arbitrary shapes in the presence of noise and does not need the number of clusters as a prior knowledge. The DBSCAN algorithm basically requires two parameters: the ϵ-neighborhood which is the minimum distance between two points and the $MinPts$ which is the minimum number of points to form a dense region. The choice of these parameters can be guided by the estimation of the number of clusters.

E. Cluster Validation

The cluster validation has been performed in terms of connectivity, Dunn index, and silhouette index.

1) Connectivity : Given a particular clustering partition $\{C = C_1, ..., C_K\}$ of the N observations into K disjoint clusters, the connectivity verifies the existence of the nearest neighbors elements in the same cluster C_K. This measure is also considered as the degree of clusters connectedness [14]. The connectivity takes its values between 0 and infinity, the minimum values are privileged.

2) Dunn Index: The Dunn index is described by [14]:

$$Dunn = \frac{\min_{1\leq i\leq j\leq K} d(C_i, C_j)}{\max_{1\leq i\leq K} |C_i|} \quad (3)$$

where $d(C_i, C_j)$ is the distance between clusters C_i and C_j and $|C_i|$ is the size of the cluster C_i. Dunn index evaluates the partitions while taking into account the distribution of objects inside classes as it is the ratio of minimum inter-cluster distance and the maximum cluster size. Larger Dunn index values are explained by a better clusters separation (high inter-cluster distances) and a compact cluster (small cluster sizes).

3) Silhouette index: The silhouette value for the i^{th} object x_i, is defined as:

$$S(x_i) = \frac{b(x_i) - a(x_i)}{\max a(x_i), b(x_i)} \quad (4)$$

where $a(x_i)$ represents the average distance between the object x_i and all objects belonging to the same cluster of x_i , $b(x_i)$ is the smallest average distance of x_i to all points in the other cluster. The silhouette value ranges from -1 to +1. A high silhouette value indicates that the sample has been well clustered, if most points have a high silhouette value, then the clustering solution is appropriate. If $s(x_i)$ is negative, the sample has been misclassified, then the clustering solution may have either too many or too few clusters. The silhouette clustering evaluation criterion can be used with any distance metric.

F. Statistical analysis

We performed a statistical analysis to examine the differences between the identified patterns using a t-test. The implementation of this statistical processing was done via SPSS 20.0 (Statistical Package for Social Sciences)1. A P-value of 0.05 was set as the criterion for statistical significance.

III. Experimental results and discussion

We implemented all aspects of knee kinematic data clustering including the non-linear dimensionality reduction using Isometric mapping, the number of clusters, the clustering using DBSCAN and cluster validation. The determined clusters are analyzed based on a clinical interpretation (Section III).

A. Estimation of the number of clusters

Fig 3 illustrates the curve of BIC and Elbow for flexion/extension (Fig 3 (a)), abduction/adduction (Fig 3 (b)), and internal/external rotation (Fig 3 (c)). In all cases, the optimal value of k is situated in the interval $[2, 4]$. Indeed, from $k = 2$, the BIC tends to change slowly and remain less changing as compared to other k's. Therefore, we limited the number of clusters to $k = 2$ for knee kinematics pattern identification.

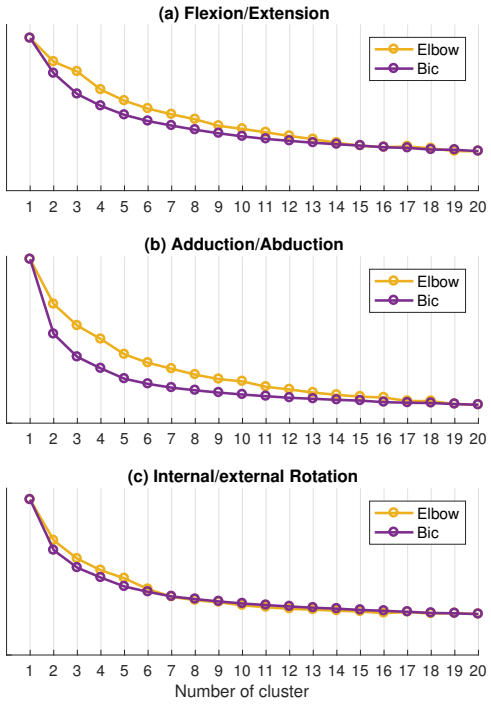

Fig. 3. Estimation of clusters number

B. Dimensionality reduction and clustering

Table I summarizes the DBSCAN parameters tuning and the cluster validation criteria, i.e., the silhouette, Dunn index, and cluster connectivity. These values show the effectiveness of the dimensionality reduction and clustering methods.

C. Knee kinematic pattern description

Once the clusters are formed, knee kinematic patterns are obtained by averaging the elements of each cluster. Fig. 4 shows the mean of each cluster describing the knee gait pattern. The analyses of the flexion/extension patterns of the Fig. 4 (a) show that the extremum amplitudes of the two clusters are observed at the same times: the maximum of the stance phase (14% of the GC), minimum of the stance phase

TABLE I
EFFECTIVENESS OF THE METHOD USING DIMENSIONALITY REDUCTION AND CLUSTERING

	Parameters tuning		Cluster validation criteria		
Planes	ϵ	$MinPts$	Connectivity	Dunn index	Silhouette
Sagittal	13	5	0	1,1e+16	1
Frontal	1	5	0	1,3e+16	1
Transverse	12	5	0	1.4e+16	1

(50% of the GC), and the maximum the swing phase (80% of the GC). However, a shift of about 10° is observed at the initial contact (1% of the gait cycle) and during the stance phase ($1\% - 60\%$ of the GC). The offset decrease during the swing phase. Statistical analysis shows that there is a significant difference between these two patterns except during the initial swing and mid swing phase (66% to 86% of the GC). Figure 4 (b) shows that the two identified patterns of the sagittal plan are different. This is confirmed by the statistical analysis which shows a significant difference during all the gait cycle. The two patterns of internal/external rotations (Fig. 4 (c)) are much more offset during the swing phase : The first pattern (Pattern 1) describes individuals a more rotated knee during the swing phase than the second pattern (Pattern 2).

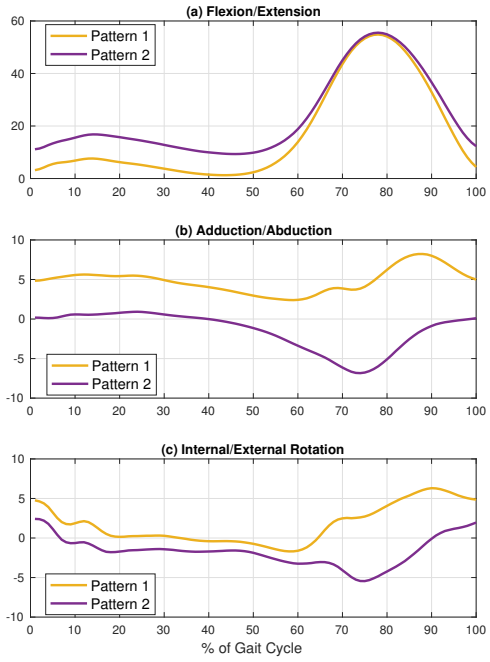

Fig. 4. Knee kinematic pattern

Moreover, we performed a gender-based analysis which shows that the men and women knee patterns are balanced between the two clusters in all of the three planes. Also, for 80% of participants, the right and left knees have been regrouped in the same cluster. This result could be of a very important clinical usefulness because, usually, in a surgical situation, the pathological knee is treated based the counter lateral knee.

IV. CONCLUSION

This study investigated nonlinear data dimensionality reduction and density-based clustering to determine knee kinematic data representative patterns of healthy knee gait. The analysis identified two representation patterns for each of the flexion/extension (sagital plane), Adduction/abduction (frontal plane) and the tibial internal/external rotation (transverse plane). Clustering quality is evaluated via general criteria, namely the silhouette width, Dunn index, and connectivity. For further understanding, the study can be extended to bi-plan and tri-plan analysis, i.e., the analysis of the combination of the abduction/adduction, flexion/extension, and internal/external kinematic data simultaneously.

ACKNOWLEDGMENT

This research was supported by Canada Research Chair on Biomedical Data Mining (950-231214). The authors would like to thank Julien Clment and Panagiota Toliopoulos for the data collection and data management.

REFERENCES

[1] J. Clement, P. Toliopoulos, N. Hagemeister, and P.-A. Vendittoli, "Healthy 3d knee kinematics during gait: differences between women and men, and correlation with x-ray alignment," *Gait and Posture*, 12 2018.

[2] N. Mezghani, A. Fuentes, N. Gaudreault, A. Mitiche, R. Aissaoui, N. Hagmeister, and J. A. De Guise, "Identification of Knee Frontal Plane Kinematic Patterns in Normal Gait By Principal Component Analysis," *Journal of Mechanics in Medicine and Biology*, vol. 13, no. 3, 2013.

[3] M. W. Whittle, *Gait analysis: an introduction*, 2007.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Ed)*. Wiley, 2001.

[5] B. Toro, C. J. Nester, and P. C. Farren, "Cluster analysis for the extraction of sagittal gait patterns in children with cerebral palsy," *Gait and Posture*, vol. 25, no. 2, pp. 157–65, 2007.

[6] G. Kienast, D. Bachmann, G. Steinwender, E.-B. Zwick, and V. Saraph, "Determination of gait patterns in children with cerebral palsy using cluster analysis," *Gait & posture*, vol. 10, no. 1, p. 57, 1999.

[7] M. J. O. Malley, S. Member, M. F. Abel, D. L. Damiano, and C. L. K. Vaughan, "Fuzzy Clustering of Children with Cerebral Palsy Based on Temporal-Distance Gait Parameters," *IEEE Transactions on Rehabilitation Engineering*, vol. 5, no. 4, pp. 300–309, 1997.

[8] N. Mezghani, M. Toumi, and A. Fuentes, "Knee Kinematic Signals Clustering for the Identification of Sagittal and Transverse Gait Patterns," in *ICCTIM 2014*, 2014, pp. 249–253.

[9] S. Lustig, R. Magnussen, L. Cheze, and P. Neyret, "The kneekg system: a review of the literature," *Knee Surg. Sport. Traumatol. Arthrosc.*, vol. 20, no. 4, pp. 633–638, 2012.

[10] J. A. Lee, A. Lendasse, and M. Verleysen, "Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis," *Neurocomputing*, vol. 57, pp. 49–76, 2004.

[11] C. Fraley and A. E. Raftery, "Bayesian regularization for normal mixture estimation and model-based clustering," *Journal of Classification*, vol. 24, no. 2, pp. 155–181, 2007.

[12] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90–95, 2013.

[13] K. Henni, O. Alata, L. Zaoui, B. Vannier, A. Alidrissi, and A. Moussa, "ClusterMPP : An Unsupervised Density-based Clustering Algorithm via Makred Point Process," *Intelligent Data Analysis*, vol. 21, no. 4, pp. 1–22, 2017.

[14] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107–145, 2001.